Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 21 (2011) 279–286

Procedia Social and Behavioral Sciences

International Conference: Spatial Thinking and Geographic Information Sciences 2011

Spatial concentrations of surnames in Great Britain

James Cheshire[a*], Paul Longley[a]

[a]*UCL Department of Geography, London, WC1E 6BT, England*

Abstract

Family names have been overlooked as a valuable source of spatially referenced population data. Presented here is a methodology published in Cheshire and Longley (2011), based on kernel density estimation that is used to identify the areas of Great Britain where any surname is most concentrated. This not only provides confirmation of a surname's geographic origin in the country but also its current spatial extent and spatial relationship with other surnames and place names. We argue that analysis using historic and contemporary data can provide baseline and change measures, and an empirical basis to change forecasting. Such analysis can provide valuable insights into national, regional and local changes in population structure, and testimony to the relevance of GIScience to population genetics, historical geography and genealogy.

Keywords: surnames; family names; kernel density estimation; Great Britain;

1. Introduction

Family names (surnames) are widely recorded in spatially-referenced population datasets. Despite their availability they have been overlooked as a source of information about population characteristics, and the long and short term dynamics that characterise population change. Presented here is a methodology designed to characterise the spatial distributions of individual surnames, and an assessment of some of the applications in population studies to which they are relevant. The material reported in this conference paper is heavily derivative of that recently published inIJGIS, see Cheshire and Longley [1].

Broadly speaking, most surnames in Great Britain are Anglo Saxon in origin, and were coined within a few centuries of the Norman Conquest in 1066, apart from those imported from abroad in comparatively recent times. Surnames are usually inherited through the male line and can provide interesting and spatially disaggregate information about their bearers, as evidenced in a small number of studies in migration, genealogy, genetics and linguistics. Most Anglo Saxon surnames can be classed as toponyms (named after specific places or geographic features), metonyms (work-based) or diminutives (‘-son’ or ‘-s’, as in Williamson or Williams). Most such names provide clues as to their geographic origins, whether this is a unique geographic location (such as the town of Rossall in Lancashire, N. England) or a broader

* Corresponding author. Tel.: +44-207-679-0500; fax: +44-207-679-0565.
E-mail address: james.cheshire@ucl.ac.uk

doi:10.1016/j.sbspro.2011.07.047

area with a particular naming convention (e.g. the '-s' diminutive suffix in Wales). Migrant surnames tend to concentrate in urban areas and, with a few exceptions (Bangladeshi surnames, for example), tend to significantly increase surname diversity in those areas.

What follows is an account of a method, outlined in more detail in Cheshire and Longley [1], capable of identifying the areas of Great Britain in which any surname is most concentrated. This not only provides confirmation of a surname's geographic origin in the country but also its current spatial extent and spatial relationship with other surnames and place names. We argue that analysis using historic and contemporary data can provide baseline and change measures, and an empirical basis to change forecasting. Such analysis can provide valuable insights into national, regional and local changes in population structure, and provide evidence of the relevance of GIScience to population genetics, historical geography and genealogy.

1.1. Surnames and geography

The contemporary spatial distribution of surnames in Great Britain is remarkably non-random [2]. Most names continue to cluster around their geographic areas of origin, providing ubiquitous cultural markers that can offer a wealth of social and cultural information [3]. In the context of historical GIS the process of identifying spatial clusters in a surname's distribution, alongside other factors such as the settlement geography or absolute size of a population, can be used to produce useful insights into historic population patterns and regional geographies.

This said, surnames have diverse origins, and their bearers have participated in different processes of economic and social change over the years, and thus the nature of geographic names concentrations is by no means uniform. Previous research into the spatial distributions of single surnames has relied on little more than visual interpretations of surname point distributions. However, the known limitations of human perception of spatial clusters argue for an automated and clearly specified parameterisation of spatial measures, which can also be used to compare different distributions in a robust and transparent manner [4]. This is illustrated by Sokal et al.'s [5] surname frequency surfaces of 100 surnames in England and Wales. In what follows, we investigate how the spatial distributions of thousands of surnames can be characterised using automated procedures, as a precursor to further spatial analysis.

In addition to a lack of standardised approaches for characterising the spatial distributions of surnames, there has been little consideration of how the underlying density of population is likely to shape observed patterns [6]. Areas of high population density are likely to have received the largest numbers of migrants of regional, national and international origin. It therefore follows that such areas will have an increased likelihood of occurrence of any particular surname [7]. In defining surname core areas, we have been mindful to avoid skewing mapped distributions towards urban areas. In addition, this research is the first to include international migrant surnames in the analysis. Such surnames identify areas that have been subject to changes in population structure. This is especially useful in urban areas that, as first destinations, have often been considered melting pots for many population groups. The boundaries between urban and rural areas are not always crisp and well defined, and may not be strictly comparable between settlements: indicators of surname diversity and extent might thus be used as criteria on which to base comparisons between rural and urban areas, and to create more meaningful distinctions between the two.

2. Data

In the analysis that follows we use the public version of the 2001 Electoral Register for Great Britain, enhanced using market research and lifestyles data by CACI Ltd. (London, UK) to improve representation of households that opt out of inclusion and non-elector groups. It contains the names and addresses of British residents aged 16 or over who are (or are about to become) eligible to vote in UK or

European elections. The surname counts were first aggregated from address level to the 218,038 2001 Census of Population Output Areas (OAs) in Great Britain using the 2001 Office for National Statistics (ONS) Postcode Directory (NSPD) (available from www.ons.gov.uk/). OAs account for, on average, 297 people in England and Wales and 199 people in Scotland OAs and, we believe, represent the most appropriate fine- scale spatial units available for analysis. The distribution of surnames in Great Britain has a very long tail – meaning that the majority of surnames are rare, but that the majority of the population does not possess a rare surname. Many low-frequency surnames, with less than 10 occurrences for example, are likely to arise from slight differences in the spelling of more common surnames, or errors in the recording process. In other cases, new low frequency surnames arise because of changes instigated by individual families (double-barrelled surnames being a common example), which render most historical inference meaningless. These records were therefore removed from the dataset.

3. Methods

The identification of geographic clusters is a well-investigated problem in spatial science, and a number of robust solutions and associated statistics are available. In this case we use kernel density estimation (KDE) which has been used in a variety of applications, including smoothing, interpolation of continuous surfaces from point data, probability distribution estimation and hotspot detection [8]. Here, KDE is used to estimate the density of occurrences of a phenomena, in this case surnames, across Great Britain. A kernel is placed over each occurrence on a regular grid. Each cell on the grid is assigned a density estimate, which is the sum of the kernel values within its locality (as defined by the bandwidth) divided by the total area of the locality from which the values are drawn.

Observed occurrences are assigned a weight according to the kernel function $k(d_{ij})$, which is a function of the distance from grid point i to observation location j. The intensity estimate at i is the sum of n individual contributions made from each observed occurrence j.

$$\hat{\lambda}_i = \sum_{j=1}^{n} k(d_{ij}) \qquad (1)$$

The extent of a kernel's influence is determined by its type. The most widely used kernel has an (unbounded) Gaussian (normal) distribution [9]. The primary effect of the use of an unbounded kernel is the production of a slightly more generalised surface because there is a less abrupt reduction in the influence of each occurrence on the surrounding grid cells. In addition to the type of kernel, the bandwidth (h) will affect the resulting density estimation. The bandwidth determines the extent of the area around each grid cell from which the occurrences, and their respective kernels are drawn. Larger bandwidths will encompass more points and therefore produce more generalised estimates than those using a smaller bandwidth. KDE therefore requires careful parameterisation in order to produce results that are representative of a surname's spatial distribution.

4. Implementation

A thorough assessment (based on a 10% sample of surnames with a range of frequencies) of the effects of different KDE parameters on a sample of our data was used to ensure the effective handling of the range of spatial characteristics exhibited by surnames (highly clustered vs. dispersed surnames, for example). It is possible to specify the bandwidth in the x and y dimensions using normal optimal smoothing for each [10]. Clearly, the spatial distribution of surnames are constrained by the coastline and therefore tend to have a greater extent in one dimension rather than the other; contrast the population distributions' of the Southwest with those along the Pennines, for example. Specifying the x and y

dimensions independently will account for this. It is also possible to specify a bandwidth that adapts to accommodate local variations in each point distribution. This can be determined by user-specified selection criteria such as the minimum number of occurrences to include within a circle centred over each occurrence [8]. Fixed bandwidths (that do not change during the KDE calculation) are generally used to represent high relative incidence within a global distribution. We found this more appropriate because variable bandwidths created additional hot-spots in towns and cities: such concentrations of population were largely absent when the names were first coined in history, and our objective was to represent high relative frequencies over more geographically extensive areas.

Finally, the frequency of surname occurrences also informs the choice of bandwidth. Rare names require much larger bandwidths relative to the density of their distributions, resulting in a greater relative spread in the KDE, in comparison with many moderate frequency surnames that attain higher densities and thus can be represented using a tightly defined KDE. In addition a point is reached where there is little gain in undertaking a KDE. The majority of rare surnames cluster in specific areas of Great Britain and automatic detection of these can be achieved using more simple methods such as a convex hull encompassing the surnames, or mapping the location quotients outlined above. This, in addition to the relative importance of chance scattering of rare name bearers, contributed to the decision to only include surnames with over 100 occurrences in our analysis.

Variations in population density play an important role in the final result. To account for this we calculate two density estimates- one weighted by the total population in each OA and the other weighted by a surname's frequency in each OA. By dividing the surname KDE by the total population KDE it is possible to reduce the spatial structure effect created by densely packed urban areas. It should be noted a balance needs to be struck when accounting for densely populated areas. Over-compensation will lead to legitimately "urban" surnames (such as toponyms, or migrant surnames) being redistributed towards less well-populated areas where they nevertheless occur in relatively low numbers. The data were spatially assigned to population-weighted centroids of each of the c.218, 038 OAs in Great Britain. The large volume of points required a balance between processing time, which increases with the grid resolution, and the level of generalisation, which has an inverse relationship with resolution. After experimentation, we found that the 16,900 cells on a 130 by 130 grid provided an acceptable compromise between the computation time for each KDE and the level of detail generated. Reducing the grid size to less than 100 by 100 began to have a detrimental impact, as the surfaces and associated contour lines drawn along them, appeared over- generalised. After the KDE was calculated the grid was clipped to the British coastline. This provided a pragmatic response to edge effects and ensured the surname distributions were plausible.

For consistency the density values assigned to each grid cell by the KDE were normalised to a value of between 0 and 1using:

$$z = \frac{a - a_{min}}{a - a_{max}} \qquad (2)$$

where*a* is a matrix of all the values on the grid. A density value of 0.95 and above therefore represents cells that fall within the top 5% of the density distribution and likewise a density value of 0.10 and below represents the cells assigned values in the lowest 10% of the density distribution.

Examples of the resulting density surfaces are provided in Figure 1. The surnames included represent the diversity of patterns produced by the KDE ranging from tightly clustered names such as Bamber, more dispersed surnames such as Palin, surnames with multiple clusters such as Khalil and those with secondary clusters in urban areas such as Macleod. KDE produces a density surface onto which a contour line can be drawn at a specified threshold value and, by straightforward extension, the density surface can be used to identify a contour that encloses a pre-specified percentage of the surname's total population. Here we focus upon population threshold contours, but also make reference to areas of highest density when considering how the contemporary distributions of surnames have spread beyond their historic core

areas to towns and cities, for example. In what follows, our density-based contours were produced using a bandwidth calculated according to the criteria suggested by Bowman &Azzalini[10] with a threshold value of 0.95 used to identify the area(s) of highest relative density of surname occurrence. This procedure was shown to indicate the shape of the cluster and to identify surnames with multiple spatial clusters.

The availability of electronic gazetteers facilitates partial validation of the areas generated for toponymic surnames. Where single or multiple matches are found between our database and the gazetteer, locational information from the latter is used to perform a point in polygon operation on the boundary of the surname core. Although a reliable indicator of the validity of the KDE procedure for toponymic names, it is not a definitive one: in many cases the place names that many surnames were derived from no longer exist, or changes in spelling of either the surname or the place name prevent any direct match with the toponym database.

5. Results and Discussion

Using the KDE methodology outlined above, core surname regions were established for 92% (27,020) of the surnames analysed. The remaining 8% did not appear to have core areas for a variety of reasons, most frequently because their extremely dispersed distributions resulted in very uniform density surfaces. Such is the number of contour lines that can be produced by this analysis it is confusing to plot them all on a single map. For ease of visualisation only the centroids of the core areas that were identified have been plotted in Figure 2 to show their distribution across Britain. The map reveals a higher density of core areas around urban centres, despite the weighting diminishing their influence in the KDE. We believe this

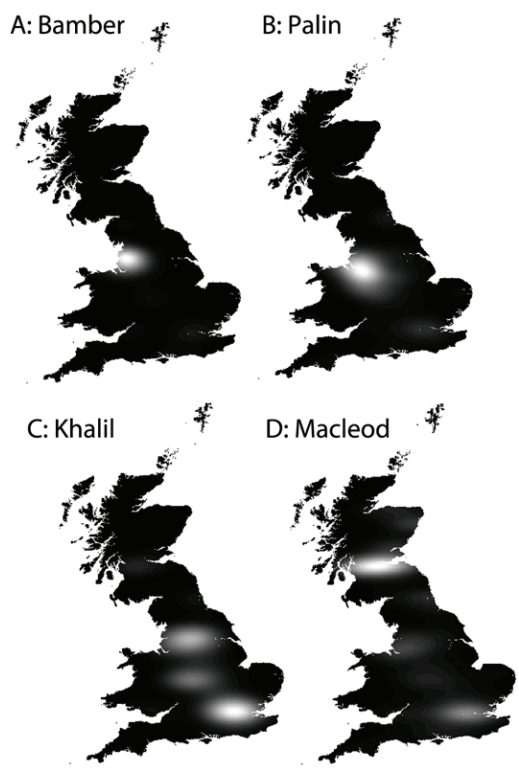

Fig. 1. Illustrative KDE surfaces for Bamber, Palin, Khalil and Macleod. Source: Cheshire and Longley [1].

to be largely an artefact of the clustering of migrant surnames. Figure 2(D) shows the distributions of place names that fall within a surname's area of origin, as defined by the 95% contour, that share the same spelling. The place names were taken from the comprehensive Ordnance Survey Placenames Gazetteer. This amounts to a validating of the toponymic origins of many surnames.

There are two outputs from the analysis for each of the 27,020 surnames: a table containing metrics and statistics; and a shapefile of the core area. The former is a database that may be queried for surnames with desired spatial attributes, while the latter provides interesting contextual information and facilitates visualisation and validation of the cores. The usefulness of this information will vary between applications.

Figure 3 shows four examples of the outputs produced by the analysis. Figure 3(A) shows the surname "Bamber". The distribution is tightly clustered in a single area of (north-west) Britain. The contour line therefore represents this closely with a tight match to the boundary of the cluster of points. This Figure also shows the cluster of points in London that are common to most surnames but that are the outcome of the sheer population size of the conurbation.

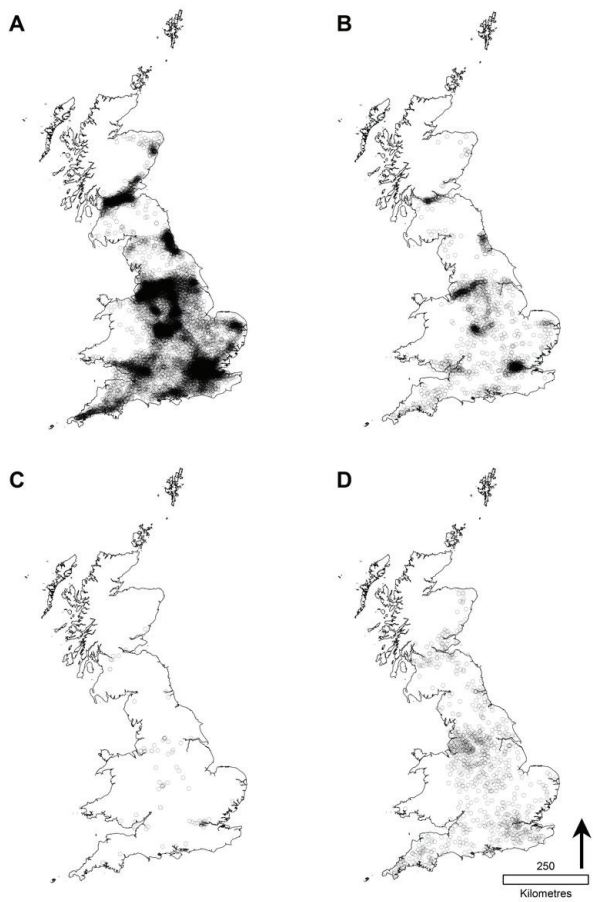

Fig. 2. The centroid locations for surnames classified as having (A) single, (B) double or (C) triple points of origin. D shows the distribution of place names that occur within the core area of a surname with the same spelling, therefore indicating a high chance of a toponymic origin for the surname. Adapted from Cheshire and Longley [1].

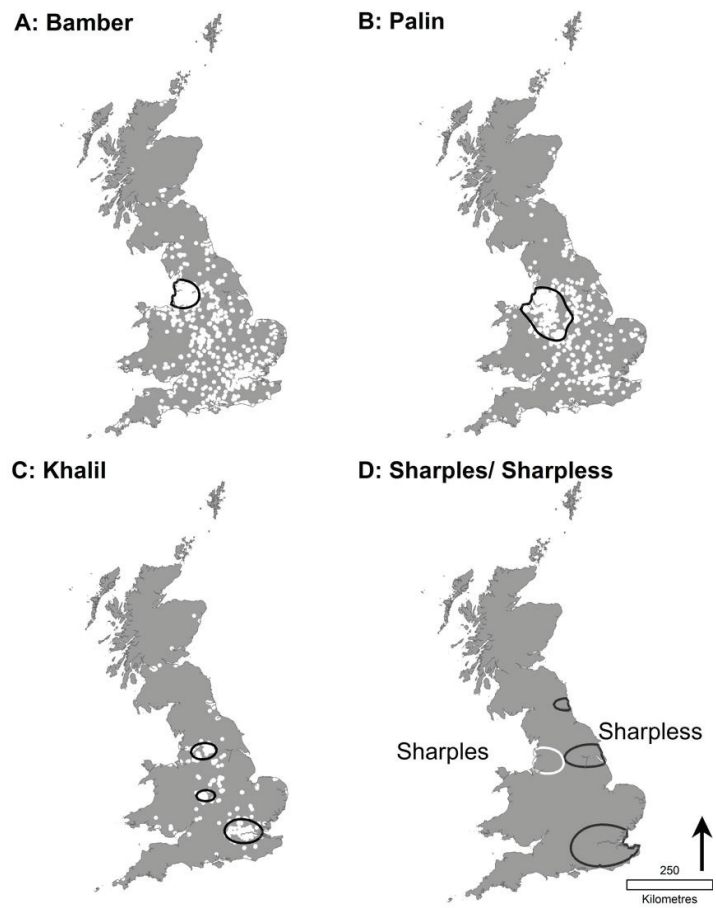

Fig. 3. Examples of surname cores with their underlying point distributions. Source: Cheshire and Longley [1].

Figure 3(B) presents the somewhat different case of the surname "Palin". Here, the point distribution is also clustered in the North West but less tightly than that of Bamber. The contour thus includes more empty space in order to capture 55% of all occurrences. This can be deduced by observing that the density of that name (frequency/ core area) within the core is only 0.10 per km^2 compared with 0.38 Bambers per km^2.

The surname "Khalil" (Figure 3(C)) demonstrates a multicentred pattern that is characteristic of many names recently imported from abroad, with tight clusters centred upon populous urban areas. This has been captured by the three contour lines. In many other cases the surname is confined to a single core around Greater London. As immigrant surnames become established, usually in London in the first instance, so their core areas typically expand into metropolitan suburbs and non-contiguous locations lower down the settlement hierarchy.

Figure 3(D) illustrates the problems of grouping surnames together that have similar spellings. Sharples and Sharpless share very similar spellings (although different pronunciations), but very different spatial distributions, with the former confined to the northwest and the latter to three areas on the eastern edge of England. Had these two surnames been combined, based on the close similarity of their spelling, the resulting spatial distribution would have been very different. It is acknowledged that many surnames were derived from the same word, or represent subtle variations in spelling from the same "root" surname. The problem, however, relates to where the line is drawn between two surnames with very similar

spellings that are unrelated to each other and those that are from the same origin but spelt differently. If two surnames are variants of a common root spelling then it is likely that they will share very similar spatial distributions, and that this will be reflected in the classification.

6. Concluding remarks

Previous attempts at classifying the spatial distributions of surnames have often been piecemeal and simplistic. Many of the datasets have been incomplete and the methods used to map unique surname distributions have been manual. We have presented an automated method that can be used with any dataset that records surnames, their frequencies, and geographic coordinates of their locations. By extension, the method presented here could be used to map the spatial diffusion of any migrant surname, or indeed any group of surnames over time. In addition, changing the parameters of the KDE make it possible for users to create a bespoke classification of a surname's spatial distribution that would closely match their purposes. Finally, the automated nature of this approach enables straightforward derivation of a different combination of metrics and the repeated application to different data sources.

This paper has sought to demonstrate the utility of intensive spatial analysis for investigating surname distributions in Great Britain. We believe it provides the best evidence yet available that the spatial origins and diffusion of a surname can be reliably captured and summarised through a series of simple metrics, No other study has undertaken this type of analysis on nearly 30,000 surnames. The preliminary nature of this study and its interest in classifying such as large volume of surnames necessitated a relatively generalised view of surnames in Great Britain. However, the greatest potential of this research may be in regional and local scale studies that seek to focus on a particular type of surname, or group of surnames. Other analysis might focus upon the dynamics of change in the core areas of particular names or groups of names, as an indicator of the opening up of regional economies. The quantitative nature of the methodology permits the easy repetition of the analysis on new datasets as they become available (such as census records) or historical data as they become digitised. The parameters of the KDEs can be amended in the light of improved substantive understanding or to meet the needs of a more focussed study. It is clear that the potential for further refinement and research is large and should be pursued to reinforce the many promising conclusions drawn here.

References

[1] Cheshire J, Longley P. Identifying Spatial Concentrations of Surnames. *International Journal of Geographic Information Science* 2011; forthcoming.

[2] Kaplan B, Lasker G. The present distribution of some English surnames derived from place names. *Human Biology* 1983; **55**(2): 243-50.

[3] Lucchetti E et al. Delimitation and aggregation between populations analyzed by surname structure. *International Journal of Anthropology* 1990; **5**(1), 49-61.

[4] Rogerson P, Yamada I. *Statistical Detection and Surveillance of Geographic Clusters*. London: CRC Press; 2009.

[5] Sokal R, Harding RM, Lasker GW, Mascie-Taylor CG. A Spatial Analysis of 100 Surnames in England and Wales. *Annals of Human Biology*, 1992; **19**:445-76.

[6] Martin D. Mapping Population Data from Zone Centroid Locations. *Transactions of the Institute of British Geographers* 1989; **14**(1): 90-7.

[7] McElduff F, Mateos P, Wade A, Cortina Borja M. What's in a name? The frequency and geographic distributions of UK surnames. *Significance* 2008; **5**(4): 189-92.

[8] de Smith M, Longley P, Goodchild M. *Geospatial Analysis: a Comprehensive Guide to Principles, Techniques and Software Tools.* 2nd ed. Leicester: Matador; 2009.

[9] Kelsall J, Diggle P. Non-parametric estimation of spatial variation of relative risk. *Statistics in Medicine* 1995; **14**(21-22): 2335-52.

[10] Bowman A, Azzalini A. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations.* Oxford: Oxford University Press; 1997.